Analysis of Valuable Clustering Techniques for Deep Web Access and Navigation

Qurat-ul-ain, Asma Sajid, Uzma Jamil
Department of Computer Science
Government College University Faisalabad
Faisalabad, Pakistan

***Abstract*—A massive amount of content is available on web but huge portion of it is still invisible. User can only access this hidden web, also called Deep web, by entering a directed query in a web search form and thus accessing the data from database which is not indexed with hyperlinks. Inability to index particular type of content and restricted storage capacity is significant factor behind the invisibleness of web content. Different clustering techniques offer a simple way to analyze large volume of non-indexed content. The major focus of research is to analyze the different clustering techniques to find more accurate and efficient method for accessing and navigating the deep web content. Analysis and comparison of Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Hierarchical and K-means method have been carried out and valuable factors for clustering in deep web have been identified.**

Keywords—Deep web; clustering; Latent Diriclet Allocation; Latent Semantic Analysis; hierarchical methods; K-means methods

I. Introduction

The complicated structure of deep web requires sophisticated methods to access and navigate the content and data on deep web databases. Unlike the indexed surface web, deep web has no hyperlink web crawling. The complexity of deep web doesn't meet up the simple navigational access methods and techniques of surface web. Thus it requires different techniques for data extraction from deep web databases.

For the enhancement of the productivity of these search engines, the programmers are trying hard to bring the content of deep web to the surface. They not only try to search valid data but also search in a way without flooding out the users with irrelevant information. Researchers and programmers of famous search engines like Google are trying to provide data which is richer in content and fulfills user demands. Google's researchers are working on algorithm for Google's Deep web crawl [17].

The main focus of research is to analyze the different clustering techniques to find more precise and better clustering technique for access and navigation of deep web. It may be truly useful to understand the minute detail of clustering techniques and algorithms and helps to put the foundation for developing more refined techniques for the data access and navigation from Deep web.

The reminder of paper is organized as follows. Section II elaborates on previous work, Section III presents the attempted dataset and proposed methodology, Section IV discusses our experimental results and the last Section V contains concluding remarks and demonstrates future work.

II. Literature Review

Various techniques of deep web clustering and classification have been presented before the comparative study of which can be found in [10]. Several researchers contributed various approaches regarding web clustering and data extraction. Here we thoroughly discuss presented clustering techniques and algorithms regarding the inspiration towards our work.

HTML structure of web documents is becoming more complex and diverse now days thus making it complicated to extract information from web pages [1]. Dr Jill Ellsworth in 1994 initially named the term "invisible web" to denote the data which was hidden from traditional search engines [2]. Google, in 2005, provides a mechanism that allows search engines and other interested users to access deep web resources and content on certain web server and database [3].

Commercial search engines have started discovering alternative method to access deep web. BrightPlanet presented the study about deep web in 2000(a massive depository of databases and data which was hidden from search engines) declaring about deep web which was 500 times greater than surface Web having indexes available at search engines [4].In deep web harvested search engines like Deep Web Harvester of Bright Planet Extract each individual word each time it access a web page [5]. U.S Naval Research Laboratory developed TOR network in 2002. Tor browser permits user to access deep web content anonymously and routing the encrypted requests so that traffic can be hidden from network surveillance tools [6].

Clustering is a nucleus task in data mining. Clustering is defined as "*the objects are clustered or grouped based on the principle of maximizing the inter-class similarity and minimizing the intra-class similarity*". Famously used clustering methods can be categorized as hierarchical methods and partitioning methods. Hierarchical decomposition can be categorized as agglomerative or divisive. Partition clustering technique generates primary partitioning and employs iterative rearrangement technique that tries to upgrade partition through proceeding clusters from a group of cluster to another. Hierarchical modeling, clustering, complex mapping and parameter knowledge gain from user connectivity based approach is need to be developed to meet the requirements [7], [8].

Clustering requires comprehensiveness, usability, ability to deal with different kinds of elements, finding of clusters with uninformed shape and ability to handle noisy data [9].

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation are widely used clustering techniques to access and navigate the content of deep web. Clustering is the partitioning of data in similar objects. Images, words, patterns and documents can be clustered. Clustering techniques for deep web pages are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Both are clustering algorithms that work on text [10].

LDA semantic based technique for heterogeneous of deep web data sources is presented. LDA is a generative probabilistic model for forming content illustration of deep web database. The document consists of topics; the core work of LDA is clustering the words in document and in topics. The DWSemClust semantics based technique is developed. CAFC-C, CAFC_CH are compared to DWsemClust. CAFC employs random document selection and CAFC_CH employs the hub based clustering of induced similarity. Both of them lack the semantic based similarity of vocabulary. DWSemClust is more suitable for sparsely distributed web sources [11], [12].

Bayesian networks are the root of many successful probabilistic topic models. But the issue of the models based on Bayesian network is the complexity of structure of model as the deduction of latent topic model distribution is frequently undetectable. LDA, hierarchical Bayesian model for the inference of topic models is much time consuming. The deep neural network (DNN) approach helps the topic model inference with low computation [21].

The classification techniques use more complex symbolic representation of instances. It does not work better on large dataset. In classification algorithm like KNN, the larger the dataset, the less accurate classification is done. Whereas Clustering techniques and algorithms are used to accelerate resource retrieval process on large dataset. It also enhances the efficiency of decision making task.

Zhengyu Yang with other fellows presented a new approach of automatic replication in SSD-HDD data processing and caching process. The exchange between Input/output performance and fault tolerance is balance efficiently through auto replica (a manager) in distributed caching and data processing systems with SSD-HDD tier storage systems [19]. Approximation algorithm is presented for automatic data placement in datacenters of all-flash multi-tier. Auto-tiering provides solution regarding allocation and migration over multiple SSD. It helps in optimizing the performance and decreasing migration operating cost. It makes the issue of polynomial time simpler and resolvable [20]. The auto-tier and auto-replica [19], [20] approach are machine learning approaches which can be further enhanced in future to operate on deep web databases.

III. Methodology

To elaborate what research is conducted it is helpful to demonstrate wholly the materials and method of research. The methodology includes the overall description of research. It provides a compact view of work which is done in this research.

- *Data Gathering*: First step was the collection of data.
- *Implementation and Comparison*: Different clustering techniques of deep web are implemented and analyzed.
- *Identification of Valuable Factors of clustering*: Valuable factors of various clustering techniques for deep web access are identified and compared. These factors include flexibility, usability, complexity, sensitivity, adaptability and scope. On the basis of these factors the comparison between differences is performed.

A. Data Collection

Data is collected from various sources. The collected data is then used for existing algorithm's implementation

1) Data Set Information (bag of words): The dataset consists of text collection in shape of bag of words. NIPS conference paper 1987 and review of Psychological articles is gathered in the dataset. After tokenization, removal of words vocabulary was diminuend by merely keeping the words coming more than 10 times. No class labels are assigned to the datasets due to copyrights factors.

TABLE I. Information about Dataset Use in Research

Data set Information				
Sr #	***Dataset***	***Source***	***Characteristics***	***Format/Pattern***
1.	Bag of words dataset (Nips and Psychreview)	UCI Machine learning Repository	Dataset Characteristics	Text
			Attribute Characteristics	Integer
	NIPS proceeding papers		Bagofwords_nips	Document word count
			Words_nips	Vocabulary
			Authors_nips	
			Authordoc_nips	Author word count
	Psych review Abstract		Bagofwords_psychreview	Document word count
			Words_psychreview	Vocabulary
2.	Temprature Sheet	National Center for Environment Information	Dataset Characteristics	Numeric
			Attribute Characteristics	Integer
3.	Movie Space	MovieLen	User movie ratings	Numeric

Table I discusses the datasets that are appropriate for topic modeling and clustering. bagofwords_nips.mat and bagofwords_psychreview file (numeric values) and authors.nips.mat and authordoc.nips.mat (vocabulary file) are provided for every text collection. The dataset is implemented on LDA and LSA clustering techniques. Document word Count means the total number of words in document.

Vocabulary in dataset means the words or letters used. Author word count includes counting of words and letter of author names.

2) Data Set Information (Temprature Sheet): National Center for Environment Information contains daily, weekly, monthly and yearly temperature forecast. The dataset includes 1981 to 2010 normal temperature consisting of 30 year temperature of all stations. The stations are represented by station numbers.

The dataset consists of daily normal weather condition and climate records. It contains most numeric values. It is very small dataset to test the efficiency of Hierarchical clustering for smaller dataset.

3) Data Set Information (movie Space): The dataset consists of user's movie rating of different years and types. The data set contains numeric values and is employed in Hierarchical Clustering, k-means clustering and pLSA.

B. Analysis and Comparision

Analysis and comparison of different clustering techniques to access and navigate the deep web are conducted to find and evaluate the functionality and performances of these techniques.

Clustering techniques are implemented in Matlab (MATLAB R2014a) for analysis. Matlab has momentous support for fast prototyping algorithms, graphing and matrix operations.

1) Techniques of clustering

a) Latent Dirichlet Allocation: Latent Dirichlet Allocation is a generative model which affirms that documents have multiple topics. Topic is a distribution over a fixed vocabulary. All documents of homogenous set contribute to the similar combination of topics but every document demonstrates these topics with distinct ratio [13].

LDA is mostly used for the modeling of text corpora. The notion of "bag of words" is implemented in these models. The topic in this model has discrete distribution of words from some finite lexicons. LDA moulds each documents as combination of clusters. Psychreview and NIPS dataset is implemented with LDA algorithm.

b) LDA Gibbs: LDA Gibbs sampling is an approach now in use to solve the good probabilities of LDA methods. Steyvers and Griffins introduced the approach of Gibbs sampling which contains the Markov Chain Monte Carlo procedure [22]. It is generally used for statistical inference. The algorithm makes use of random numbers and produces different results when executed. Execution of LDA Gibbs (basic Topic Model dataset of psychreview and nips' bagofwords is performed to extort the set of topics and presents the most liable words per topic.

Algorithm

1. Input: bag of words (consisting of number of times each word occur)
2. Output: Topic assignment to each word token
3. Calculate the number of times each word is given the topic.
4. Number of times the topic is allocated to document.

c) Latent Semantic Analysis: LSA attempts to map words and documents in concept space or clusters for comparing by implementing centroid-based clustering. It compares the meanings and concepts behind the words. It analyzes and examines the documents for finding original concepts and notions of these documents [14], [15]. Some suitable conditions to apply LSA techniques are as follows:

1) When documents contain same writing style.

2) When each and every document has focuses on particular topic.

3) When a word has higher probability of belonging to a topic than another topic and lower probability with other topics.

d) pLSA: Probabilistic Latent Semantic Analysis is an upgrade to LSA technique. Words in topics from pLSA are closely related than words in LSA. Topics are multinomial random variables in pLSA, and a particular topic produce each word and thus various words are originated by various topics. The larger the number of documents the larger the pLSA model, is the limitation of pLSA model.

Table II highlights the differences between pLSA and LSA. The LSA method originates from Linear Algebra and acts upon the Singular Value Decomposition (SVD). The pLSA method has the foundation on mixture decomposition. The advantage of using pLSA statistical model over SVD is that it permits to join diverse models methodologically.

TABLE II. DIFFERENCE BETWEEN LSA AND PLSA

Sr#	***Latent Semantic Analysis***	***Probabilistic Latent Semantic Analysis***
1	Highest Gaussian Error	Highest Likelihood Function
2	No apparent explanation of parameters	Polynomial Distribution of Parameters
3	Singular Value Decomposition is precise.	pLSA EM congregate to confined best possible

2) *Types of clustering*

a) Hierarchical Clustering: Hierarchical algorithmic methods use similarity or distance matrix. Splitting or merging of one cluster is performed at one time. Dendrograms are used to represent hierarchical clustering.

Hierarchical Methods

Divisive: Divisive is the top bottom approach for clustering. Divisive clustering is less blind to the global structure of data. The idea of divisive clustering is that all objects are in one cluster. The cluster is divided into sun-clusters which are sequentially separated into more sub-clusters. This process persists unless the preferred cluster is acquired. The divisive clustering follows the top-down approach of hierarchical structure.

Agglomerative: Every object embodies its own cluster. The clusters are sequentially merged unless the desired pattern of cluster is achieved. The fundamental function of agglomerative clustering is the calculation of proximity among two groups of clusters. Agglomerative clustering follows the bottom up hierarchy [16].

1) Initiate with point as single clusters

2) At every step, merge the closest pair of clusters until only a cluster left.

Algorithm

1. Calculate the proximity/similarity matrix
2. Let each data point be a cluster
3. Merge the two nearest and most similar closest clusters. Update the proximity/similarity matrix
4. Repeat 3 & 4 until all patterns are in individual Cluster.

b) K-means Clustering: K-means is a heuristic approach of partitioning clustering. Each cluster is connected with a central point called centroid. Each point in cluster is linked to cluster with nearest and closest cluster. Number of clusters must be identified and denoted as *k*. The aim is to reduce the summation of distances of the points to their relevant centroid. Mixture model (EM algorithm: dealing with clusters having uncertainty), k-medoids(better for noise and outlier), k-median and k-models are the variations of k-means method.

Algorithm

1. Choose *K* points as early centroids.(initial centroids are selected randomly)
2. From *K* clusters allocate all points to the nearest and closest centroids.
3. Recomputed the centroid of every cluster
4. Reiterate step 2& 3 unless the centroids don't change.
5. Selection of K points may be performed by using some method.

IV. Results and Discussion

Results are presented in this section by implementing various algorithm and different dataset. Major outcome of this research is describes below.

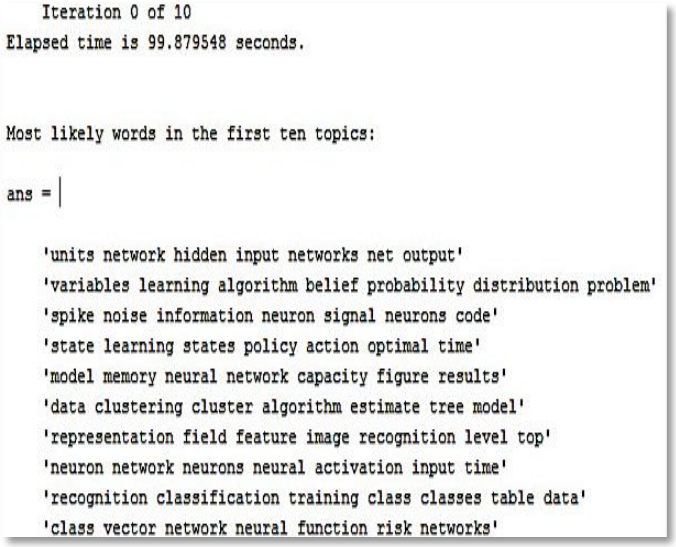
A. *Latent Dirichlet Allocation*

LDA Gibbs algorithm is implemented on the bag of words (NIPS and psychreview) dataset. It proves better than LSA traditional algorithm by accessing and extracting desired information. It is proved as time efficient techniques with a large dataset. As the number of iterations increase the time efficiency of LDA is disturbed. Words extraction from different topic models is depicted as below. Table III shows the detailed comparison of LDA and LSA.

Fig. 1 shows the most occurred words in first ten topics with ten iterations. It offers more compact view of data extraction from documents. LDA (LDA-Gibbs) technique is more accurate to present possible desired results.

TABLE III. Comparision between LDA and LSA

Comparision between LDA and LSALatent Dirichilet Allocation (LDA) vs. Latent Semantic Analysis (LSA)		
Factors	**Latent Dirichilet Allocation (LDA)**	**Latent Semantic Analysis (LSA)**
Usability	More effective at finding word-level topics with large dataset.	Less effective as compared to LDA
Time Complexity	Less Time consuming	Much Time consuming
Suitability	Suitable for large dataset as well as smaller data set	It performs efficiently with smaller data set but it is not suitable for large dataset
Flexibility	Gibbs LDA sampling is easier to compute	Singular Value Decomposition is difficult to compute
Capacity	Provide a probabilistic model at document level	Offers no probabilistic model at document level
Usage	It assigns Probability for document/topic/word in each cluster	Probabilistic LSA defines the probability of /topic/word in each cluster.



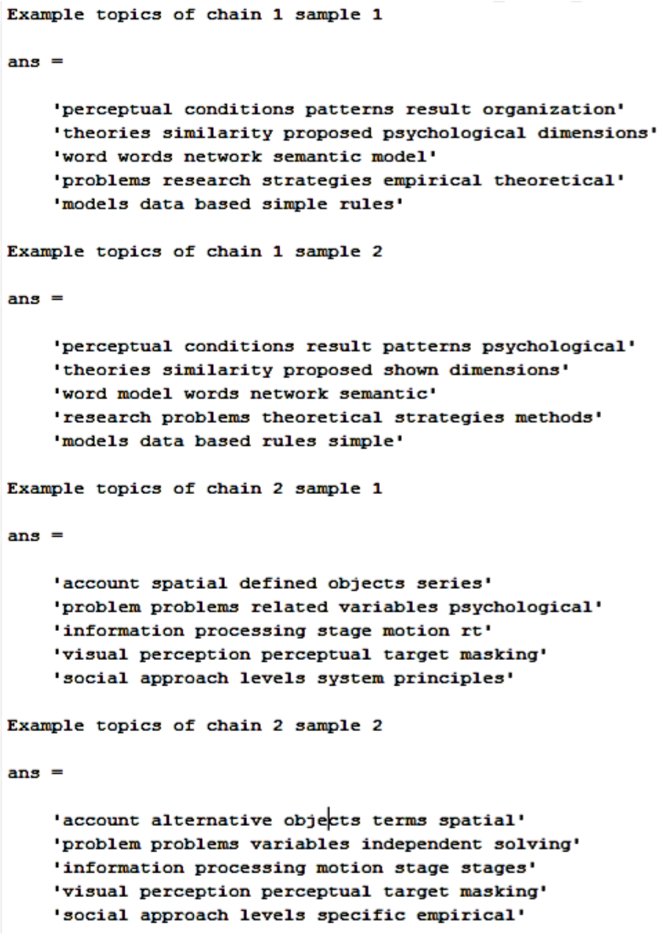
```
    Iteration 0 of 10
Elapsed time is 99.879548 seconds.


Most likely words in the first ten topics:

ans =


    'units network hidden input networks net output'
    'variables learning algorithm belief probability distribution problem'
    'spike noise information neuron signal neurons code'
    'state learning states policy action optimal time'
    'model memory neural network capacity figure results'
    'data clustering cluster algorithm estimate tree model'
    'representation field feature image recognition level top'
    'neuron network neurons neural activation input time'
    'recognition classification training class classes table data'
    'class vector network neural function risk networks'
```

Fig. 1. LDA Word extraction in different Topics.



```
Example topics of chain 1 sample 1

ans =

    'perceptual conditions patterns result organization'
    'theories similarity proposed psychological dimensions'
    'word words network semantic model'
    'problems research strategies empirical theoretical'
    'models data based simple rules'

Example topics of chain 1 sample 2

ans =

    'perceptual conditions result patterns psychological'
    'theories similarity proposed shown dimensions'
    'word model words network semantic'
    'research problems theoretical strategies methods'
    'models data based rules simple'

Example topics of chain 2 sample 1

ans =

    'account spatial defined objects series'
    'problem problems related variables psychological'
    'information processing stage motion rt'
    'visual perception perceptual target masking'
    'social approach levels system principles'

Example topics of chain 2 sample 2

ans =

    'account alternative objects terms spatial'
    'problem problems variables independent solving'
    'information processing motion stage stages'
    'visual perception perceptual target masking'
    'social approach levels specific empirical'
```

Fig. 2. LDA topic generation.

Fig. 2 depicts that Topis generated by LDA algorithm in 100 iterations from 2 samples. The lesser the iterations the lesser the time consumption is observed.

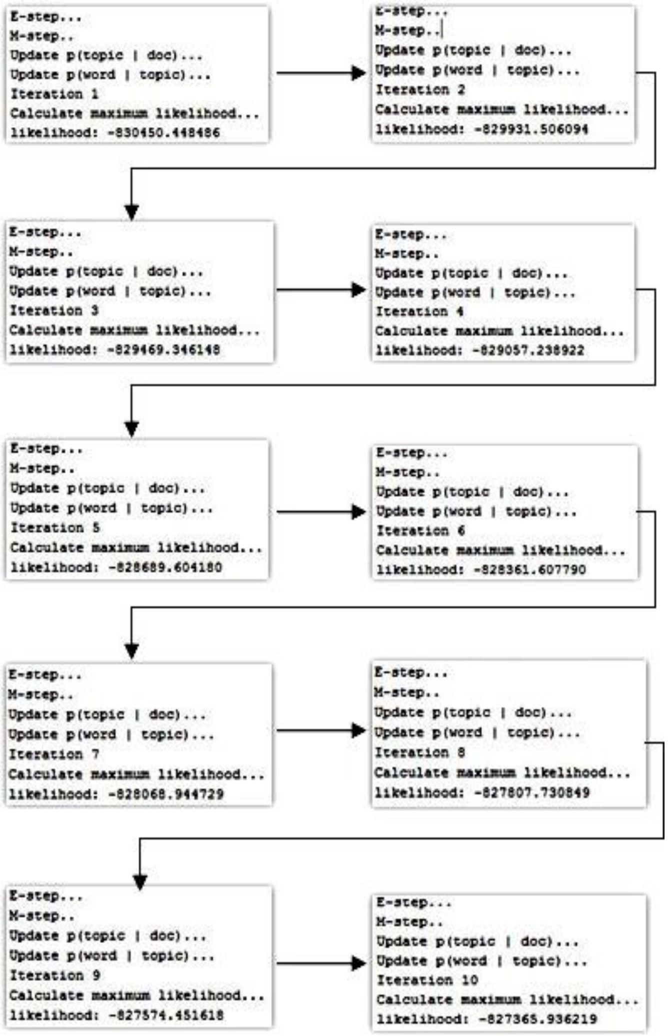
B. Latent Semantic Analysis

LSA algorithm is implemented on Bag of Word (NIPS & psychreview) dataset. The LSA measures the likelihood of every word in a topic model. The word extraction from topics is a time consuming process in LSA as compared to LDA. A small dataset is easily accessed in lesser time than large dataset.

A pLSA code with a large data set with different iterations has run and produced different elapsed time. The execution time on large dataset with 100 iterations is 2331.868618 and 121.118317 sec with 50 iterations. Table III compares LDA and LSA on the basis of various factors.

The pLSA working of algorithm with minimum iteration produce fast execution, whether the execution on large dataset with much iteration is time-consuming.

Fig. 3 shows the likelihood of occurrences of words in topics through pLSA EM (Expectation-Maximization) steps with 10 iterations. It generates results with top 20 words in top 10 topics.



```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 1
Calculate maximum likelihood...
likelihood: -830450.448486
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 2
Calculate maximum likelihood...
likelihood: -829931.506094
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 3
Calculate maximum likelihood...
likelihood: -829469.346148
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 4
Calculate maximum likelihood...
likelihood: -829057.238922
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 5
Calculate maximum likelihood...
likelihood: -828689.604180
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 6
Calculate maximum likelihood...
likelihood: -828361.607790
```

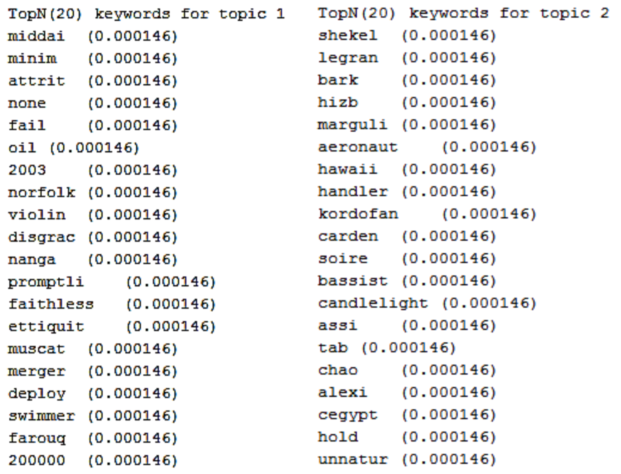
```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 7
Calculate maximum likelihood...
likelihood: -828068.944729
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 8
Calculate maximum likelihood...
likelihood: -827807.730849
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 9
Calculate maximum likelihood...
likelihood: -827574.451618
```

```
E-step...
M-step..
Update p(topic | doc)...
Update p(word | topic)...
Iteration 10
Calculate maximum likelihood...
likelihood: -827365.936219
```

Fig. 3. Likelihood of occurrences of words in a topic.



```
TopN(20) keywords for topic 1    TopN(20) keywords for topic 2
middai   (0.000146)              shekel   (0.000146)
minim    (0.000146)              legran   (0.000146)
attrit   (0.000146)              bark     (0.000146)
none     (0.000146)              hizb     (0.000146)
fail     (0.000146)              marguli  (0.000146)
oil (0.000146)                   aeronaut     (0.000146)
2003     (0.000146)              hawaii   (0.000146)
norfolk  (0.000146)              handler  (0.000146)
violin   (0.000146)              kordofan     (0.000146)
disgrac  (0.000146)              carden   (0.000146)
nanga    (0.000146)              soire    (0.000146)
promptli     (0.000146)          bassist  (0.000146)
faithless    (0.000146)          candlelight (0.000146)
ettiquit     (0.000146)          assi     (0.000146)
muscat   (0.000146)              tab (0.000146)
merger   (0.000146)              chao     (0.000146)
deploy   (0.000146)              alexi    (0.000146)
swimmer  (0.000146)              cegypt   (0.000146)
farouq   (0.000146)              hold     (0.000146)
200000   (0.000146)              unnatur  (0.000146)
```

Fig. 4. pLSA's words extraction from topics with possible likelihood.

Fig. 4 shows pLSA topic extraction with word's occurrence likelihood that these words and keywords are extracted from each topic with maximum likelihood of word occurrence in these topics. The 10 iterations create 10 topics with each topic consisting of 20 top keywords in those topics.

C. Hierarchical clustering

Hierarchical clustering algorithm is implemented on Movies and Temperature datasets [18].

1) Types of Linkage Function

a) **Single Linkage**: Merging of two clusters where two nearest elements have the minimum distance. It produces the minimum spanning tree. It promotes the expansion of extended clusters. It is highly sensitive to the noise.

b) **Complete Linkage**: Merging of two clusters in every step that merging has the maximum distance. It promotes dense clusters. It doesn't work efficiently if extended clusters are presented.

c) **Average Linkage**: Keeping in view the sensitivity of complete linkage clustering to outliers and the predisposition of single linkage clustering to create big chains that don't match up the discerning idea of clusters as solid objects is observed. Agglomerative clustering is very strong and vigorous with average cluster distance and linkage. Fig. 5 shows the comparisons of types of linkages.

The algorithm is implemented on MovieSpace dataset and Temperature Sheet dataset. As the hierarchical clustering is implemented on large dataset of MovieSpace, it consumes more CPU time and memory space and affects the cost of Input/output. It is slower than k-means algorithm on same dataset and takes longer time to generate result.

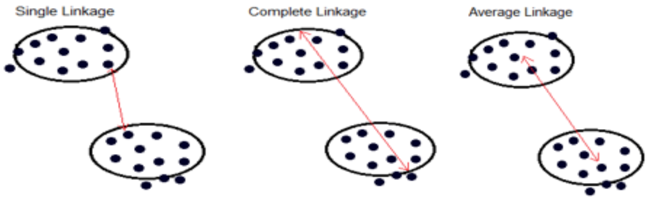


Fig. 5. Single, complete and average linkages.

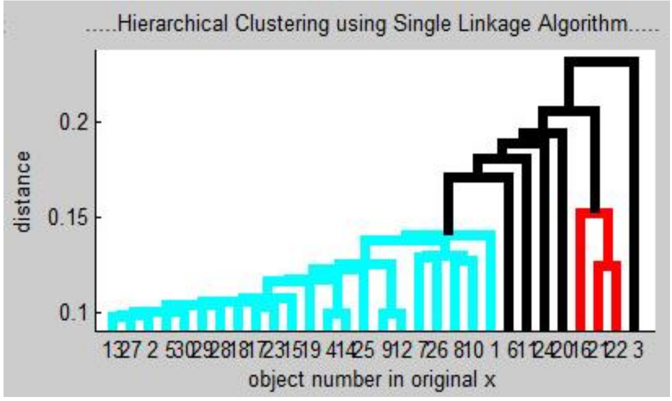


Fig. 6. Hierarchical clustering using single linkage algorithm on large dataset.

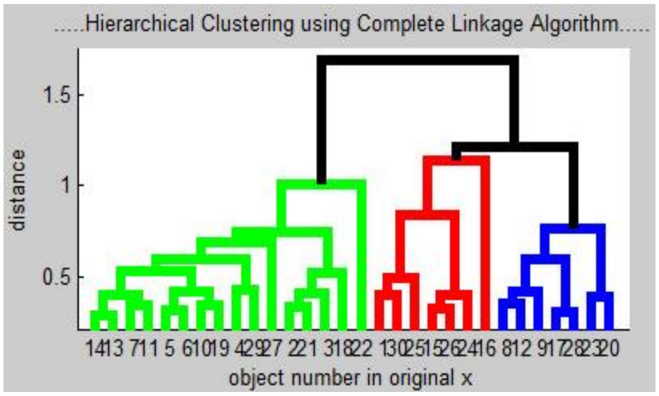


Fig. 7. Hierarchical clustering using average linkage algorithm on large dataset.

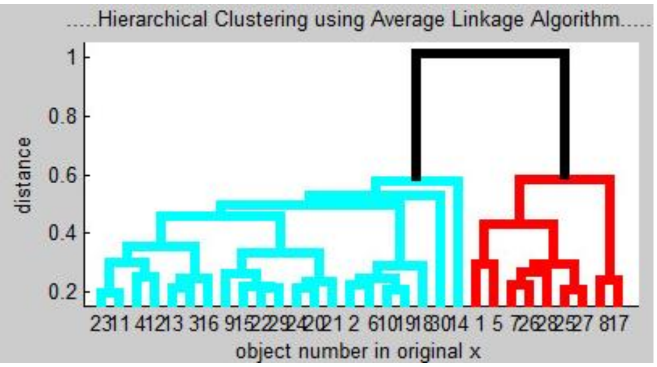


Fig. 8. Hierarchical clustering using complete linkage algorithm on large dataset.

The implementation of hierarchical algorithm on smaller dataset of Temperature sheet produces different results. The small number of instances in dataset results in low Input/output cost and less execution time. It produces following results on movieSpace dataset.

In Fig. 6, 7 and 8 the pairs of object forming cluster are depicted in object number in original X (Y label). These figures show the hierarchical tree of Single, complete and average linkage function which was performed for hierarchical clustering on MovieSpace dataset. After analysis of these figures, the execution time for the movieSpace dataset was obtained is 282 sec.

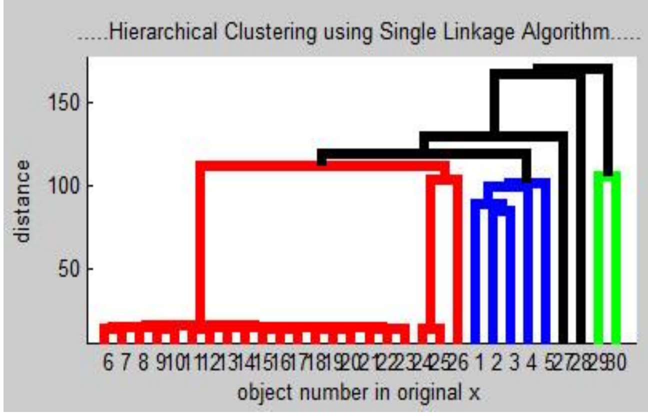


Fig. 9. Hierarchical clustering using single linkage algorithm on small dataset.

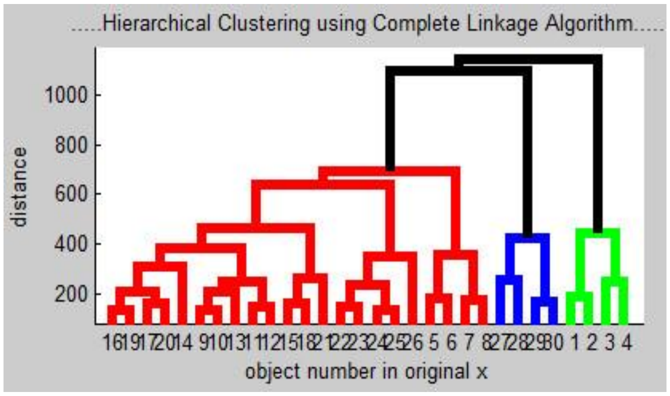


Fig. 10. Hierarchical clustering using average linkage algorithm on small dataset.

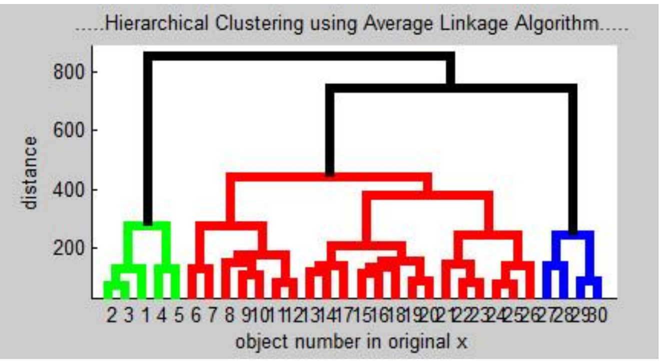


Fig. 11. Hierarchical clustering using complete linkage algorithm on small dataset.

Fig. 9, 10 and 11 shows single complete and average linkage algorithm implementation for small dataset of temperature respectively. Execution time for Temperature dataset is 41 sec.

D. K-means Clustering

K-means algorithm is implemented on MovieSpace dataset. On data set of MovieSpace the K-means algorithm is implemented which produce faster results and low execution time than the same dataset's implementation in Hierarchical algorithm. The algorithm produces random cluster each time when executed. It generates efficient result with sparse (not dense) data with no noise. It utilizes less Input/output and memory storage for execution. K-means performs better with large data set as compared to hierarchical clustering.

Table IV displays the complete comparison of hierarchical and K-means clustering.

TABLE IV. COMPARISION BETWEEN HIERARCHICAL CLUSTERING AND K-MEANS CLUSTERING

Hierarchical Clustering vs. K-means Clustering		
Factors	***Hierarchical Clustering***	***K-means Clustering***
Nesting	Combination of nested clusters, which are arranged as tree.	Combination of objects in related clusters such that every object is in just a single cluster.
Complexity	Time and Space complexity(Non-Linear) Time complexity is at least *O(m^2)* m is the total number of instances that is not linear with number of objects in cluster. Clustering a large dataset may have immense I/O cost.	Linear Complexity The algorithm works well with very large number of instances. Better of Large dataset.
Sensitivity	Problem with noise and outliers in data	Problem with noise and data that has outliers. Appropriate only when mean is characterized. It needs the number of clusters in advance.
Result Demonstration	The result of hierarchical clustering is presented in the form of dendrogram.	The results of K-means clustering are presented mostly in cluster points and plots.
Back-tracking	No back-tracking is observed as hierarchical clustering can never go back to previous step	K-means is a randomized algorithm , it always select clusters randomly each time
Usability	The Use of Hierarchical clustering is normally constrained with numeric attributes. A hierarchy of documents in deep web database can also be maintained	The Use of K-means is frequently constrained with numeric attributes.
Adaptability	Show better performance on data set consisting of non-identical clusters containing string and having a common centers or clusters	Show better performance on data set which has isotropic clusters and not as adaptable as hierarchal single link method.

TABLE V. K-MEANS IMPLEMENTATION OVER DATASET

Sr#	*K*	*D*	*N*	*X*	*INIT*	*answer*
1.	2	2	3000	5	19	(3 2 3 1)
2	3	3	3000	5	19	(1 2 1 3)

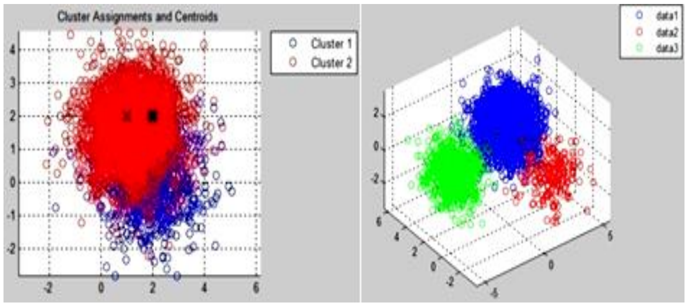


Fig. 12. Views of clusters assignment and centroids.

In Table V, K-means takes required number of clusters and the starting means as inputs and generates final means as output. Means of cluster are described first and last means. K is the number of clusters, *d* is dimensions (2nd dimensional or 3rd dimensional), X and INIT are the required numbers of clusters and initial means.

Fig. 12, Sensitivity of K-means can be seen in center initialization of a cluster. When the center initialization is done poorly it may lead to bad intersection speed and on the whole bad clustering. K-means clustering method groups the same the type of objects in similar cluster.

V. SUMMARY AND CONCLUSION

The complicated structure of deep web requires sophisticated methods to access and navigate the content and data on deep web databases. The comparative analysis of clustering techniques demonstrates that to extract information from deep web databases is the complex task. It deducts the weaknesses of these techniques to overcome for better performance. LSA is beneficial to work with small datasets; it is much time-consuming working with large datasets. The LDA (LDA-Gibbs) technique is far better than LSA to present possible desired results. Hierarchical and partitioning methods are beneficial for structuring the content on deep web databases. The random allocation of documents in cluster having similar or dissimilar documents produces time efficient method, whereas, hierarchical structure of documents of similar category produces time consuming methods. The combination of both methods may enhance some features for structuring document.

The analysis provides new directions to refine these techniques. The future work focuses on designing, modification and amalgamation of existing techniques for better performance and functionality. Genetic algorithm based clustering techniques are taken into consideration for future technological enhancements. The Combination of clustering technique with other data mining or machine learning technique like Deep Neural networks and artificial neural networks may provide a more optimized and refined technique of data access on deep web. For maximum desired search outputs semantic as well as syntactic accuracy of search prediction can be devised through discovering unique techniques to deal with certain semantic and linguistic properties of deep web sites.

REFERENCES

[1] Lavanya M. & Dr. Usha Rani "A framework for vision-based Deep Web Data Extraction for web " , September 2012.

[2] Michael K. Bergram "The Deep Web: Surfacing Hidden values". White Paper: *Deep Content* , September 2001.

[3] Chertoff M. & Simon T.(2015,February)The Impact of the Dark Web on Internet Governance and cyber security. *Paper Series No 6.Global Commission on Internet Governance*

[4] Steve Pederoson(2013,March) Understanding the Deep Web in 10 Minutes. White Paper.

[5] BrightPlanet (2012), What is Deep Web Harvest? https://brightplanet.com/2012/07/what-is-a-deep-web-harvest/

[6] Dinglein R. , Mathewson N. & Syerson P.(2004,August) "Tor: Second Generation Onion Router". *In proceeding of 13th conference on USENIX Security Symposium.* Volume 13.

[7] Han J. et al.,"Data Mining: concepts and techniques". *Third Edition,* 2012

[8] Wensheng Wu et al.,(2004,June) "An Interactive Clustering-based Approach to Integrating Source Query Interfaces on the Deep Web". *In proceedings of SIGMOD June 2004.*

[9] Estivill Castro V.,(2002, June) "Why so many clustering algorithm- A position Paper". *SIGKDD.* Volume 4,Issue 1.

[10] Muhunthaadithya C & Rohit J.V et al., "Clustering of Deep WebPages: A comparative study".*International Journal of Computer Science Iformation Technology(IJCSIT)* .Volume 7, No 5, October 2015

[11] Umara Noor & Ali Daud et al., "Latent Dirichlet Allocation based Semantic Clustering of Heterogeneous Deep Web Sources", September 2013 [*In proceedings of 5th International Conference on Intelligent Networking and collaborative Systems*]

[12] Ben Eysenbach(2016,May) Latent Dirichlet Allocation and Application to DSPACE.

[13] David M.Blei , Andrew Y.Ng et al., (2003, March) Latent Dirichlet Allocation. *Journal of Machine Learning Research*

[14] Katzman D. (2010,June) "Clusters that Think". *Article. Link* http://deepwebtechblog.com/clusters-that-think/

[15] Landauer T. K. & Peter W. Foltz et al., (1998) "An Introduction to Latent Semantic Analysis". *Discourse Processes,* 25,259-284

[16] Manpreet Kuar and Usvir Kuar."Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection". *International Journal of Advanced Research in Computer Science and software Engineering.* Volume 3,Issue 7,July 2013

[17] Jayant Madhavan, David Ko et al,. "Google's Deep-Web Crawl"

[18] Sivtalana Volkova, "Latent Dirichlet Allocation".*Final Year Project Report*"Hierarchical clustering Algorithm" https://github.com/meskatjahan/Hierarchical-clustering-Algorithm

[19] Zhengyu Yang et al., "AutoReplica: Automatic data replica manager in distributed caching and data processing systems".

[20] Zhengyu Yang et al., "AutoTiering: Automatic Data Placement Manager inMulti-Tier All-Flash Datacenter"

[21] Dongxu Zhang etal., "Learning from LDA using Deep NeuralNetworks"

[22] "LDA Gibbs". *Topictolbox* https://github.com/huashiyiqike/topictoolbox